Numerical Weather Prediction

Report on calibrating the site-specific probability forecasts from the ECMWF Ensemble Prediction System

Forecasting Research Technical Report No. 386

James Denholm-Price and Ken Mylne

email: nwp_publications@metoffice.com

©Crown Copyright

Report on calibrating the site-specific probability forecasts from the ECMWF Ensemble Prediction System.

Dr James Denholm-Price and Ken Mylne

June 26, 2002

The Previn Development Project was intended to improve the use of the Ensemble Prediction System (EPS) from the European Centre for Medium-Range Weather Forecasting (ECMWF) as currently implemented in Previn (which stands for 'Predictability Visualisation'). Here we describe the part of the project designed to improve the reliability of the site-specific probability forecasts derived from the EPS data (The original Previn system was uncalibrated.) Calibration is performed in two stages. Firstly rank histograms from past verification are used to re-weight ensemble members, giving improved probabilities for event thresholds within the spread of ensemble members. Secondly to calibrate thresholds lying outside the spread of the ensemble, Weibull distributions are fitted to past ensemble outliers. This report describes the calibration processes, the significance testing of fitted distributions and demonstrates the significant improvements in probability forecasts achieved.

1 Project background and rationale

Previn produces probabilistic information from EPS fields interpolated to specific sites. Verification rank histograms from the original Previn forecasts (Mylne et al. 2001) have overpopulated extreme ranks which suggest that the ensemble has insufficient spread, although this may not be the only reason for overpopulated outliers (*e.g.* see Hamill (2001)). Non-flat rank histograms correspond with unreliable probability forecasts, as illustrated by verification reliability diagrams (described below). Mylne et al. (2001) showed that there was a significant mismatch between issued probabilities for a particular event and its frequency of occurrence in the original Previn system and this may be improved by calibration.

2 Probabilistic Calibration

In this section the calibration procedure is described. This is discussed in some detail as the method adopted is different from the more common calibration method based on the reliability diagram (which is also summarised below).

The original Previn system produced probabilistic forecasts of *local* weather parameters (12 hour precipitation accumulations, 10m wind speed and 2m temperatures) at 42 UK sites using the 51

member EPS. These forecasts were routinely verified against observations taken from the sites, results of which may be found in Legg et al. (2001). The forecasts and their verification are influenced by deficiencies in the EPS, by the difficulty of relating grid-box based NWP output to local weather and any errors in the observations. The new Previn system arising from the Development Project in 1999–2001 operates at many more sites worldwide (although only 30 UK sites were used during its development) and in addition to the above parameters, also produces forecasts of 24h. precipitation and daily maximum and minimum temperatures. Temperatures, including daily max/min, and wind speeds are derived from a Kalman filter, taking inputs from several model predictor fields, which is used to correct site-specific biases. Precipitation is simply bias-corrected. Woolcock (2001) describes the Kalman filter and the development of the probabilistic calibration system used to improve the reliability of the forecast probabilities is described here.

During the development of the calibration system a 60-day running mean bias correction was applied to archived EPS data to approximate the effect of the Kalman filter. Thus in the discussion below, all EPS data are either bias-corrected (1998-9 data) or Kalman-filtered (2000–1). Additionally the EPS horizontal resolution was upgraded from T_L159 (approximately 120km over the UK) to T_L255 (80km). Together these affect the statistics used for calibration, an effect that will be eliminated during 2002 when the system no-longer relies on data from before 2001.

2.1 Reliability

Categorical forecasts are the most common types of forecast issued in probabilistic terms. For example, "what is the probability that 2m temperature will fall below freezing in 5 days' time?" might be translated (for our purposes) to "$P(T < 0°C)$ at T+120h". The correspondence of P with event frequency is called *reliability*. A reliability diagram (*e.g.* see Wilks 1995, p266) starts by subdividing such categorical probabilistic forecasts (drawn from many verified forecasts) into probability bins, say 0–0.05, 0.05–0.15, ..., 0.85–0.95, 0.95–1. The relative frequency of the forecast event occurring, conditional upon it being issued with forecast probability lying within the range of a particular bin, is compared with the forecast probability (*i.e.* the average probability for the bin). Ideally these should be equal so that, for example, an event forecast with 70% probability occurs 70% of the time when verified over many cases. Plotting forecast probability against verifying frequency should ideally produce a diagonal line joining $(0, 0)$ to $(1, 1)$, or for the above probability bins the line would join $(0.025, 0.025)$ to $(0.975, 0.975)$.

Calibrating specific categorical probability forecasts may be performed using the reliability diagram to remap forecast probabilities to their 'reliable' verification frequencies. For example if a particular forecast when issued at 70% probability verifies with a relative frequency of 60% then the 'reliable' forecast to issue would be 60% instead of 70%. Using this method, each forecast threshold, forecast variable and season must be calibrated separately and these must all be repeated whenever the ensemble prediction system changes, which is time and resource intensive.

2.1.1 Calibration *via* rank histograms

Following Hamill and Colucci (1997) a more general form of calibration has been performed, based on the verification rank histogram (or just 'rank histogram', discussed below). In this case the information needed for calibrating probabilities is calculated just once for each variable, irrespective of the forecast threshold. The statistics still need to be updated to account for seasonal variations and model changes but the use of one set of calibration statistics for any forecast at a given time is efficient and convenient.

The verification rank histogram gives the frequencies of the verification occurring in each of $N+1$ bins formed by sorting the N members of the ensemble into numerical order (ranking). Each pair of ranked members defines the edges of a bin, giving $N-1$ 'closed' bins. Additionally there are two so-called *outlier* bins with one 'open' end and whose right/left (upper/lower) edges are the largest/smallest ensemble members, respectively. If the ensemble members are denoted by $e_1 \ldots e_N$ then the relationship between the ensemble members and bins is illustrated by the following:

	e_1	e_2	e_{N-1}	e_N	
lower outlier bin 1	bin 2	$\cdots$	bin N		upper outlier bin $N+1$

An example of such a rank histogram from the 51 member EPS is shown in figure 1 below. An ideal ensemble system should produce a rank histogram with equal frequency in each of the bins (Anderson 1996), *including* the outlier bins which are open on one side, *i.e.* the verification should lie above or below the ensemble with the same frequency as any other bin, which is 1/52 in the EPS.

The rank histogram relative frequencies are estimates of the probability that a given verifying observation will fall into a given ensemble bin. Departures from an ideal 'flat' histogram indicate an increased or decreased likelihood of the verification occupying a given bin. A higher rank histogram frequency than 1/52 in a given bin implies an excess of occasions when the verification falls in that bin (and *vice versa*). For example, if the ensemble has insufficient spread then the verification will too often fall outside the spread of the ensemble, increasing the population of the extreme ranks and decreasing the population of the inner ranks.

2.1.2 Categorical probability forecasts from NWP ensembles

To assign a probability to a categorical forecast (*e.g.* $P(T < 0°\mathrm{C})$) from an ensemble, traditionally we count the number of members of the ensemble with $T < 0°\mathrm{C}$ and divide that by the total number of ensemble members to give a probability. This is equivalent to adding 1/51 for every ensemble member that passes the threshold, *i.e.* they all have equal *weight* (1/51).

This approach fails if the ensemble does not encompass the chosen threshold, *e.g.* if in this case all ensemble members have temperatures greater than 0°C. Strictly the above 'count the members' method would yield a probability of zero even though $T = 0°\mathrm{C}$ may lie close to the lower extreme of the (ranked) ensemble. By analogy with the rank histogram discussion above, it is more realistic to interpret the ensemble probability in terms of bins. In this case the

probability of the temperature being less than all members of a perfectly calibrated ensemble is 1/52 (**NB** not 1/51 as we now consider the probability to be spread over $N+1$ bins delimited by the N ensemble members). Similarly, the probability of the temperature being greater than every ensemble member is also 1/52, as is the probability of temperatures lying inside any bin delimited by two distinct ranked ensemble members (for an ideal ensemble). So if we denote the 51 ensemble members by $e_1 \cdots e_{51}$ and the weight assigned to each bin by $w_1 \cdots w_{52}$ we have for a weather parameter X and chosen threshold x:

$$P(X \leq x) = \begin{cases} w_1 & x = e_1 \\ \sum_{j=1}^{i} w_j & x = e_i \quad i = 1 \ldots 51 \end{cases} \tag{1}$$

$$P(X > e_{51}) = w_{52} \tag{2}$$

where $\sum_{i=1}^{52} w_i = 1$.

Thresholds which do not coincide with an ensemble member can be treated in an obvious way. The probability weight assigned to that bin is linearly interpolated from the lower member up to the upper member. This is equivalent to linearly interpolating between points in the cumulative distribution that is defined by the ensemble members. It is not immediately clear how to treat thresholds outside the range of the ensemble. There *is* a probabilistic weight assigned to the whole 'bin'. The weight is either w_1 for the *lower outlier bin* $(\infty, e_1]$ or w_{52} for the *upper outlier bin* (e_{52}, ∞) – the 'open' and 'closed' parentheses arise from the convention adopted here that $w_1 = P(X \leq e_1)$. A method for assigning probabilities within the outliers is discussed below.

For an ideal ensemble the weights assigned to each bin are identical but must be specified for any practical ensemble which suffers from, for example, a lack of spread. Correctly weighting the bins *calibrates* the probabilities derived from the ensemble. Estimates of the weights can be found by verifying a large number of forecasts to evaluate the verification rank histogram.

2.1.3 Probabilistic weights

Figure 1 shows a rank histogram derived from the Kalman filtered EPS temperature field showing relative frequencies of 2m temperature observations from UK sites for T+12 hours. The outlier bins are coloured blue and red to denote observations where the *ensemble* temperature was colder (blue) or warmer (red) than the observations. Evidently the outliers are significantly overpopulated relative to the 'ideal' (horizontal line) – note that the vertical scale is logarithmic. This situation is typical of the EPS when verified against observations as the Kalman filter is designed to correct site-specific biases and not the ensemble spread.

Towards the end of the 10 day forecast the ensemble spread is better and there are fewer outliers so the rank histogram is (sometimes) flatter – this varies between parameters and depends on the EPS and Kalman filter performance. An example of a better rank histogram is shown in figure 2 for 2m temp. at T+240h. The outlier bins are much smaller and the interior bins correspondingly larger at T+240h than at T+12h (fig. 1). Note, however that at T+240h (fig.

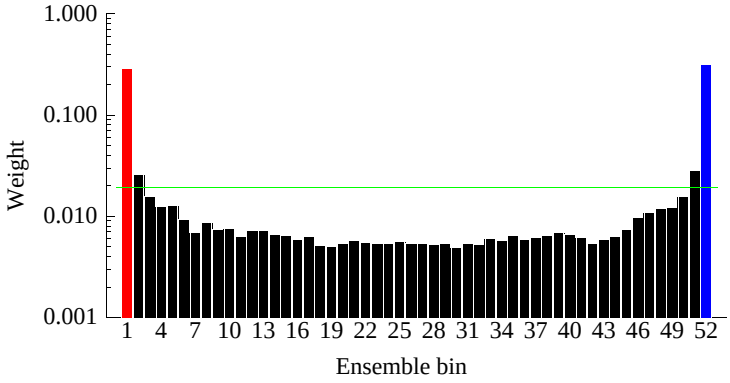

Figure 1: Rank histogram from Kalman-filtered EPS 2m temperatures at T+12h using data from Winter 2000/2001 (DJF).

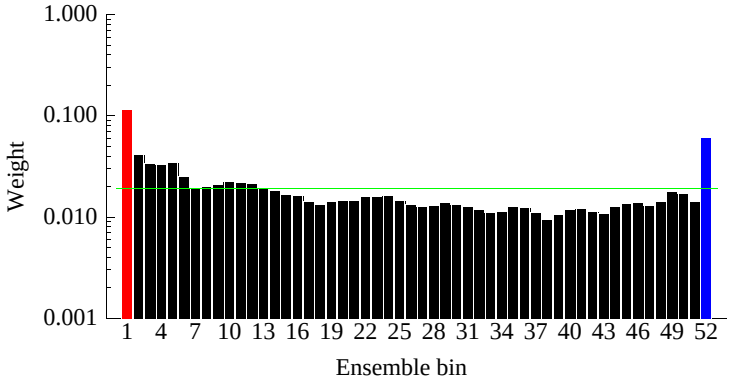

Figure 2: Rank histogram from Kalman-filtered EPS 2m temperatures at T+240h using data from Winter 2000/2001 (DJF).

2) there is a slope from left to right in the interior which may indicate a bias in the forecast temperatures (see Hamill (2001) for more discussion of this).

Calibrating the ensemble using its verification rank histogram assigns larger weight to bins in which relatively more verification events are found than would be expected, and *vice versa*. This has the effect of *increasing* the forecast probability within the bins with increased weight and *vice versa*. This increases the reliability of the forecast by more accurately relating the forecast probabilities to the observed frequencies.

3 Outlier probabilities

Figures 1 and 2 show that the outlier bins are often overpopulated when using the EPS to forecast local weather. The weights assigned to these bins can be very large – at short forecast ranges, up to 30 times larger than an interior bin. The EPS gives no information about the shape of the distribution within the outlier bin and when weights are large, linear extrapolation

from the extreme ensemble members is likely to be inaccurate. However it was found that the distribution of the difference between the observation and the nearest ensemble outlier (either the upper or lower extreme members) often exhibits a shape that can be modelled by a Weibull distribution whose probability density function (PDF) is given by equation 3 below. Hamill and Colucci (1997) used a Gumbel distribution with rank-histogram based calibration for short-range quantitative precipitation forecasting . Gumbel and Weibull are both examples of extreme value distributions and are related to the more familiar Gamma distribution (*e.g.* see Wilks 1995).

$$\mathcal{P}(x;\alpha,\beta) = \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] \tag{3}$$

Here α and β are related to the *location* of the maximum and *width* of the distribution, respectively. The Weibull cumulative distribution function (CDF) is:

$$\mathcal{C}(x;\alpha,\beta) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right] \tag{4}$$

Fitting $\mathcal{C}$ to the verifying observations lying outside the ensemble (in the upper or lower outliers) parametrises the shape of the upper or lower outlier tail. Given α and β for a given ensemble outlier bin, eq. 4 gives probabilities conditional upon the threshold x being within that bin. This can be used to evaluate probabilities for *any* threshold within the outliers (*i.e.* any threshold which happens to lie outside the ensemble). Probabilities from the ensemble can then be calculated using the following equation:

$$P(X \le x) \;=\; \begin{cases} w_1\left(1-\mathcal{C}(e_1-x;\alpha_l,\beta_l)\right) & x \le e_1 \quad \text{lower outlier} \\ \sum_{j=1}^{i-1} w_j + w_i\frac{x-e_{i-1}}{e_i-e_{i-1}} & e_{i-1} < x \le e_i \\ & 2 \le i \le 51 \quad \text{ensemble interior} \\ \sum_{j=1}^{51} w_j + w_{52}\mathcal{C}(x-e_{51};\alpha_u,\beta_u) & x > e_{51} \quad \text{upper outlier} \end{cases} \tag{5}$$

where (α_l,β_l) and (α_u,β_u) refer to fitted Weibull distributions parameters from the *lower* and *upper* outliers, respectively. As before, the weights sum to unity $\left(\sum_{i=1}^{52} w_i = 1\right)$ and the linear interpolation within interior bins has been written explicitly for $e_1 < x < e_{51}$.

3.1 Fitted Weibull distributions

To find α and β the Weibull CDF, eq. 4, was fitted to the estimated CDF from each outlier distribution using a nonlinear least-squares routine. Examples of the outlier distributions (shown as their histograms which are estimates of the outlier PDF) together with the fitted Weibull PDF are shown in figures 3 – 5 with data taken from Winter 1998/9 (DJF). These were chosen to be fairly typical of the range of distributions found in the 1998/9 bias-corrected dataset. Similar shapes were found with the Kalman-filtered data, although the length of the tail is often reduced and Weibull fit improved by the Kalman filter corrections and/or improvements to the EPS in the 2001 data, which is explored further below. Therefore the 1998/9 data are a more extreme test of the use of Weibull tails – if the procedure works with the bias-corrected 1998/9 data then it should work well when combined operationally with the Kalman filter.

In order to accumulate sufficient data from which to derive reliable statistics it is necessary to group stations together. In principle the choice of membership of the 'collections' should be made to separate stations for which Kalman-filtered EPS data have different distributions. This is not known *a priori* and would require more data than are available for each version of the EPS (prior to upgrades). Pragmatically, the grouped stations should be chosen sufficiently far apart that their statistics are approximately independent but within the same 'geographical region' so as to remain identically-distributed. Fortuitously the original Previn 42 station set and the reduced set of 30 stations used during development are all within the UK yet none is particularly close to another.

In each of figures 3 – 5 the caption includes the fitted Weibull parameters α and β together with the number of data points N that contribute to the histogram. The data are the outliers of the Kalman-filtered ensemble, which varies between stations. Each station may not contribute an observation to the histogram from every verification so the number N varies.

The data are made *approximately* independent by enforcing at least a 24h. separation in time between points from the same station. It is recognised that the data are not truly independent and identically distributed. A compromise has been necessary so that there are sufficient data available to build a representative histogram without drawing data from so many months that either the EPS has changed (*e.g.* in November 2000 the EPS resolution was upgraded) or that data were taken from months that were unrepresentative of the current season. These compromises are discussed further in section 4 where the operational implementation of the calibration system is considered.

3.1.1 2m temperature

Figures 3(a) illustrates a good fit of the Weibull PDF to the outlier distribution (for the lower outlier 2m temperature at T+12h) and similar shapes are found at all forecast ranges (T+12–240h.) Figure 3(b) (upper outliers at T+240h.) illustrates the worst situation where the outlier distribution is clearly bimodal and the fit is poor. In the latter case the second mode in the large tail is largely due to the combination of data from 'coastal' and 'inland' stations in the collection of stations included in the 1998/9 dataset. Subjectively separating the stations near the coast from those well inland (not shown) reveals that the outliers from the coastal stations are clearly bimodally-distributed whilst the reduced set of inland stations is much closer to the Weibull distribution (*e.g.* the chi-squared value for the inland stations is 3.0 which would not be rejected even at the 99.9% confidence level whilst for the coastal stations it is 16.2 which would be rejected at the 80% level). This indicates that the bimodal character which causes the poor Weibull fit in fig. 3 is due to some characteristics of the stations used (most likely their location).

The Kalman filter is designed to correct for site-specific variations by generating a local bias correction from the past 60 days' verification. The filter uses weather parameters in the filter that are not restricted to the variable being corrected. For example, including wind direction when correcting for 2m temperature allows the filter to correct for direction-dependent biases such as might be observed for a coastal site (see Woolcock 2001 for a full discussion). The combination of increased EPS resolution (upgraded from T_L159 to T_L255 in December 2000) and the Kalman filter seems to be effective in reducing such site-specific errors that lead to

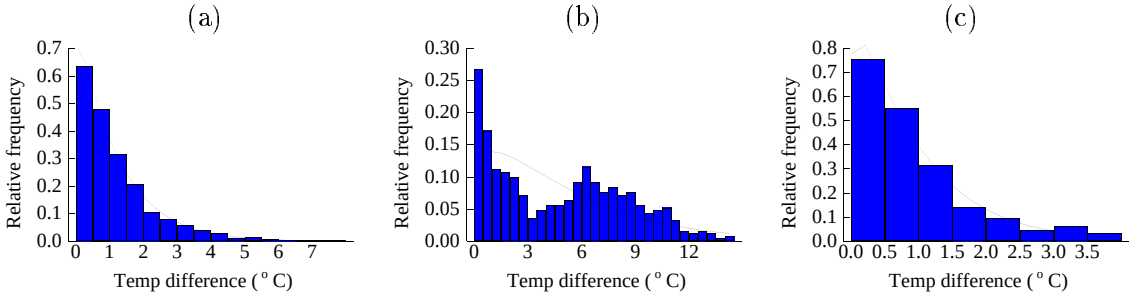

Figure 3: 2m temperature outlier histograms using winter (DJF) data: (a) Lower outliers, 1998/9, T+12h. ($\alpha = 1.0$ and $\beta = 1.2$, $N = 834$) (b) Upper outliers, 1998/9, T+240h. ($\alpha = 1.2$ and $\beta = 5.5$, $N = 500$) (c) Upper outliers, 2000/1, T+240h. ($\alpha = 1.1$ and $\beta = 1.0$, $N = 127$)

bimodal outlier distributions. Figure 3(c) shows the histogram and fitted Weibull curve for Winter 2000/1 using the Kalman-filtered EPS data. Comparing with figure 3(b) it is clear that not only is the bimodal character much weaker in Winter 2000/1 but also that the tail length (maximum distance of outlying observations from the most extreme ensemble member) is smaller. As far as a single verifying season can be taken as being representative, this suggests that the higher resolution EPS together with the Kalman filter effectively combine to give outlier histograms that are more easily characterised by the Weibull distribution.

3.1.2 10m wind speed

Figures 4(a,b) show wind speed upper outlier distributions also from Winter 1998/9. Fig. 4(a) illustrates the good correspondence between the outlier distribution and fitted Weibull at T+12h, which is also observed in the majority of lower outlier distributions of wind speed. As wind speed is constrained to be positive the lower outlier distributions tend to become narrower at longer forecast ranges (as the EPS spread increases more members predict lower winds). However the spread does not increase so much that the outlier distribution vanishes and the Weibull fit continues to be good.

The upper wind speed outlier distribution in fig. 4(b) fitted illustrates a different problem with fitting a Weibull distribution. In this case there is a broad maximum at the origin, which is made up of the outliers closest to the EPS member with largest wind speed. Whilst the tail of the distribution is well described by the fitted Weibull, for $\alpha > 1$ the PDF has a minimum at the origin which does not fit the outliers' behaviour but is due to the relatively broad peak of wind speed differences between 0 and 5ms^{-1}. Fortunately the Kalman filter effectively narrows the peak (as well as shortening the tail so that the Weibull distribution fit is closer, as shown in fig. 4(c), again bearing in mind that data from only one season have been used in the latter figure.

3.1.3 Precipitation

Finally in fig. 5 three upper outlier distributions of 12h. precipitation (PPT) accumulations are shown. These data exhibit a comparatively long tail which the Weibull distribution does

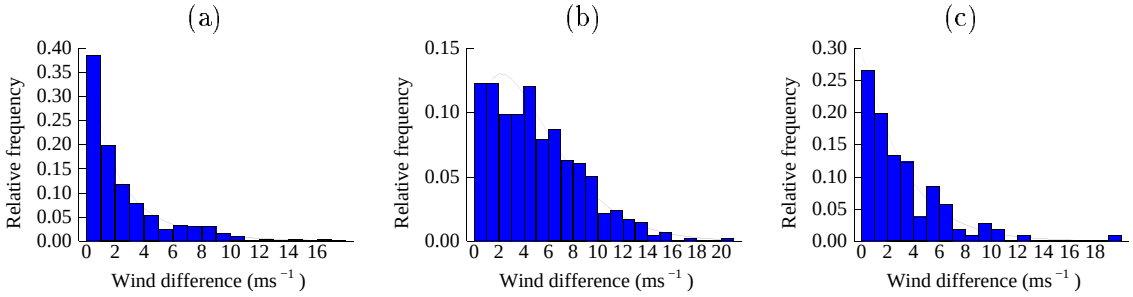

Figure 4: 10m wind speed. Upper outlier histograms (Winter) (a) 1998/9, T+12h. ($\alpha = 0.8$ and $\beta = 2.4$, $N = 594$) (b) 1998/9, T+240h. ($\alpha = 1.4$ and $\beta = 5.6$, $N = 414$) (c) 2000/1, T+240h. ($\alpha = 1.0$ and $\beta = 3.3$, $N = 117$)

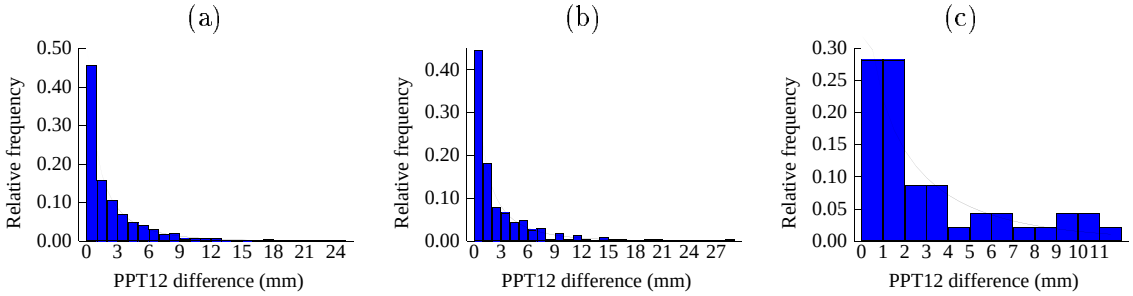

Figure 5: 12h. PPT accumulation. Upper outlier histograms (Winter) (a) 1998/9, T+12h. ($\alpha = 0.7$ and $\beta = 2.1$, $N = 989$) (b) 1998/9, T+240h. ($\alpha = 0.7$ and $\beta = 2.2$, $N = 226$) (c) 2000/1, T+240h. ($\alpha = 0.8$ and $\beta = 3.0$, $N = 48$)

not reproduce accurately (the fit concentrates at the more highly-populated left-hand end in fig. 5). However the situation is improved somewhat by the Kalman filter in figure 5(c) from Winter 2000/1, where a shorter tail is evident. The small amount of data ($N = 48$) makes this assessment less reliable, but similar behaviour is observed at all forecast ranges.

The lower outlier distributions for precipitation are not shown as they are dominated by a single bin at the origin. This is caused by the dominance of zero or low precipitation events (and forecasts) together with the constraint that PPT > 0. As such there is no need for a fitted 'tail' at the lower end of the PPT PDF and it is sufficient to calibrate the lower PPT outliers using the weights alone.

4 Operational calibration system

The calibration and fitted Weibull tails discussed above are used in the updated Previn system. The facility exists to produce probability forecasts either from the raw EPS, Kalman filtered (subsequently denoted 'KFMOS'), calibrated using the weights alone ('CAL') or calibrated with weights and the Weibull tails ('CAL+W'). Calibration PDFs are available for 2m temperature, 12 and 24 hour accumulations of precipitation (separately), wind speed and maximum/minimum daily temperatures, which are derived from the Kalman filter rather than from EPS direct model

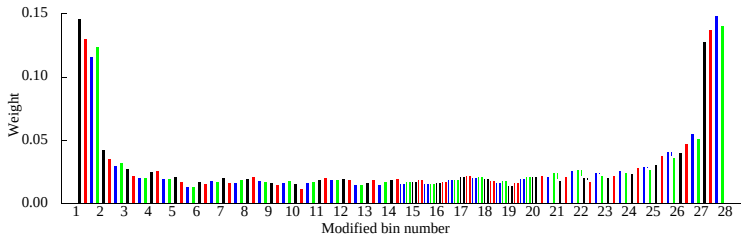

Figure 6: Comparing weights – black is 'ideal' (see text); red, blue and green are taken from 1, 2 and 3 month averages, respectively.

output. For precipitation (as explained above) only the upper outlier has a Weibull tail fitted to it. Otherwise PDFs for each calibrated variable can be used with (or without) both upper and lower tails.

The Kalman filter updates the EPS each day, based on the past 60 days' data which allows for a reasonably quick response to changes in the EPS as well as inter-seasonal variability. The calibration (weights) and Weibull tails (α and β) also respond to such changes but a longer cycle is needed to accumulate sufficient data for the calibration. The rank histograms and outliers are recalculated at the start of each month, which balances the need to respond to EPS changes which may occur on time scales of much less than one season against the computational cost. The amount of data needed to generate reliable calibration statistics is now considered.

4.1 Stability of weights

The probabilistic weights needed to calibrate EPS forecast PDFs must be estimated from past data. They evolve over seasonal time scales, especially the outlier weights which have a proportional effect on the inner weights (since the weights must sum to 1). To assess how much data are needed to generate approximately statistically stable weights, figure 6 shows a histogram of the weights for 2m temperature valid at T+60h. from April 1999 using different averaging periods.

The 'truth' in this case is difficult to assess – one month's data shows significant variability within the inner bins (where one would expect a smooth variation over a sufficiently long average). Instead the black bars in fig. 6 represent the truth for April as an average of the rank histograms from March, April and May (1999). Note also that as plotted the 52 bins have been reduced for clarity by averaging neighbouring pairs of interior bins (excluding the 2 outliers) to reduce the number of bins 50+2 to 25+2 for each month. This has a small smoothing effect but any significance tests were performed on the complete data.

Figure 6 is fairly typical of the weights from all variables and forecast times as far as the variation between periods is concerned. There is evidently some variability between 1, 2 and 3 month averages. Significance tests with the Kolmogorov-Smirnov (using the asymptotic 95% limit of Erickson and Taylor 1989), and Chi-squared statistic between each averaging period (red, green and blue bars) and the 'truth' (black) reveal that only the one month average fails to be similar to the truth at the 99% (chi-squared) or 95% (K-S) significance level. Similar tests pooling data

from both 1998 and 1999 reveal a similar level of significance. Therefore in order to preserve this stability during periods of significant changes (such as the transition from statistics based on the T_L159 to the T_L255 resolution EPS) a three month running average was adopted in the operational system so that the weights for forecasts in the coming month are based on rank histograms from the preceding three months. The monthly evolution of the outlier weights using this procedure is displayed operationally – an example is shown in figure 10 below for the weight assigned to the upper outlier bin for 2m temperature forecasts at all forecast ranges (T+0, 6, ..., 240).

4.2 Resampling the Weibull fit

The Weibull tails are fits to the distributions of verifying observations that fall outside the EPS. Since the size of the outlier weights decreases with forecast range (as the EPS spread increases) it is necessary to include data from a longer verification period to get a good sample. This is exacerbated by the need to use independent data which, as noted above, is approximated by including outliers from individual stations separated by 24h. Tests have shown that for the experimental set of 42 UK stations, 6 months' data are sufficient to populate the tails sufficiently at all forecast ranges. To avoid spanning seasons the Weibull tails use data from the current (forecast) year and data from the previous year. Initially the data were taken from the last 3 months 'this' year and 3 months 'last' year starting from the current month, *e.g.* the calibration for November 2001 would use August, September and October (ASO) 2001 data together with SON 1999 data – 2000 data were not archived whilst the Kalman filter was being setup. This had the disadvantage that the calibration throughout 2001 relied partially upon data that were 2 years old, coming from the T_L159 EPS without the benefit of the Kalman filter. In 2002 this will no-longer be necessary.

In winter 2001 the strategy for choosing months was changed after the verification of max/min temperature forecasts was found to be much worse than that for spring 2001, due in part to the over-reliance on data from the past season. Instead a more 'centred' approach was adopted, so that November 2001 calibration took data from ASO2001 and OND1999. Unfortunately this need for large sampling times means the Weibull tails will respond more slowly to changes in the EPS than the weights, which are calculated using data from just 3 months. Hopefully the climatology of the outliers does not change too rapidly (something which remains unproven as its test would require a much larger archive of EPS verification than is available).

To facilitate operational monitoring of the tails a *resampling* (Wilks, 1995; pp. 145–148) procedure was adopted. The data populating the outlier bins is pooled (with imposed 24h. separation) and the global fit to the outlier histogram determined (α_g and β_g). The 'goodness of fit' of the Weibull distribution to the outlier histogram can be assessed by such measures as the chi-squared (χ^2) and Kolmogorov-Smirnov 'D' test (which is also known as the Lilliefors test when testing a fit against data not independent of the fitting procedure). Here D is defined as the maximum absolute distance between the Weibull fit and the outlier histogram so that small values of D correspond with good Weibull fits.

From the pool of outliers a random sample of at most 200 points is chosen with replacement (so individual points may be repeated within the sample) and the curve-fitting procedure repeated on this subsample. By repeatedly sampling-with-replacement (1000 times) a distribution of α

and β is formed. Provided that the initial pool of data is sufficiently large and is representative of the outliers' climatological distribution then the bootstrapped distribution of α, β and D show how 'representative' the global fit is of that climatology. The bootstrapped distributions enable the calculation of confidence intervals for α_g and β_g which give an idea of how reliable the fitted tail is likely to be in the chosen forecast period (up to 1 month ahead).

Bootstrapped distributions of α, β and D are usually slightly skewed versions of a Gaussian shape. Examples are shown in figures 7(a–c), each of which comes from the calibration for June 2001 (using data from March/April/May 2001 and April/May/June 1999). In the operational system these figures are generated for every variable (2m level/max./min. temperature, 12 and 24h. PPT accumulations and 10m wind speed), outlier bin (upper/lower) and forecast range (T+0–240h. in 6h. steps). These provide a means of monitoring the behaviour of the calibration system *via* the bootstrapped distributions.

Figure 7(d) combines the data from figs. 7(a) and (b) into a scatter plot of α against β and this illustrates the shape of their joint distribution. The ellipse is a 90% confidence contour (see Press et al. 1992, §14.5), the dashed lines univariate limits and the points are colour-coded according to their acceptance/rejection level from the standard χ^2 test on each resampled Weibull fit. This figure allows a rapid overview of the resampling procedure and can highlight potential problems. For example a large numbers of red points would indicate a very poor set of resampled curve fits (50% χ^2 confidence or less) and a non-oval scatter might suggest that outliers were not well represented by the global Weibull fit. Figure 7(d) also highlights a compromise made in the operational system: Ideally the resampling of the Weibull tails should be repeated so that the oval is more densely populated (say 10,000 times) but it is restricted to 1000 resamples by the time taken to process the data.

4.3 Operational constraints and calibration problems

Graphical summaries of the calibration statistics are also produced from the operational system. Examples of these may be found in figures 8(a) and (b), 9 and 10 which show, respectively: Time series' of α_g and β_g (with error bars, shown staggered about the forecast range for clarity), percentiles from χ^2 and D statistics and outlier weights are plotted from calibration runs from four consecutive months.

Figs. 8(a) and (b) illustrate how the Weibull tails evolve over the forecast range and month-by-month. The outlier tail (β) is often longer with larger spread at later forecast ranges. There is also an oscillatory 'diurnal' cycle in these upper temperature outliers, with longer tail lengths at mid-day (T+0 is 1200 UTC), suggesting the outliers are furthest from the ensemble during the day. This is confirmed by the upper outlier weights shown in figure 10 which take their maximum values at mid-day verification times (T+0, 24, ..., 240h.)

The unusual behaviour of the March (black) calibration at 12h. intervals in figs. 8(a) and (b) arises because 6 hourly data were only available from December 2000. Thus the archived outlier data included in the pool from 1999 contribute *no* points to the calibration at T+0 or T+6, 18, 30, 42, ... which makes the March calibration less reliable at times (large error bars). This behaviour improves in later months but the problem will only be eliminated in March 2002 when the use of 1999 data can be discontinued.

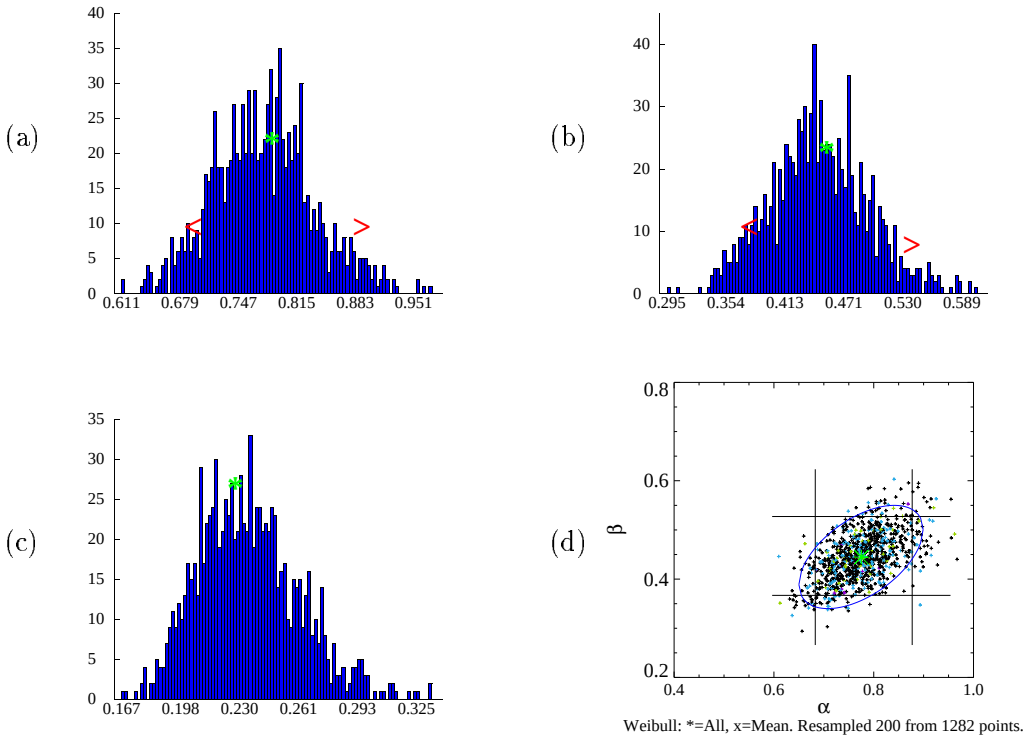

Figure 7: Bootstrapped distributions of (a) α, (b) β, (c) D and (d) α and β as a scatter plot. In each figure '*' (green) shows the position of the global fit. In (a–c) the 90% confidence intervals on α_g and β_g are '<' and '>' (red) in the histograms as are the vertical and horizontal dashed lines in the scatter plot. In (d) the ellipse also contains 90% of the points and each point is colour-coded according to its χ^2 'goodness of fit' (black, blue, and green for 100-95%, 90–95% and 75–90%, respectively).

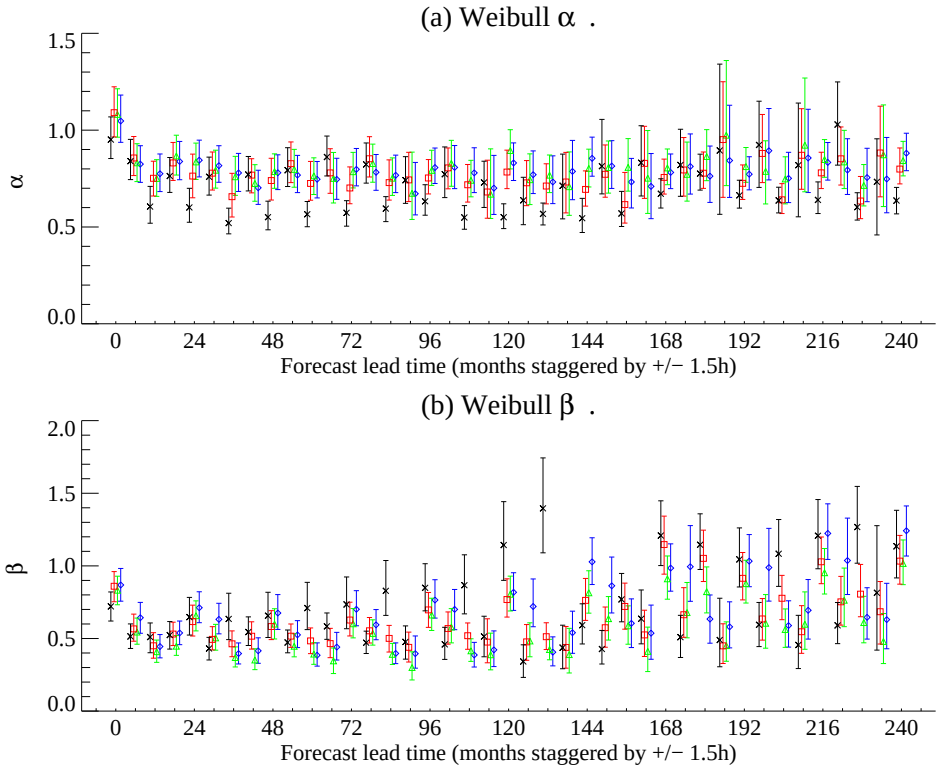

Figure 8: Weibull parameters from the quasi-operational calibration system for upper 2m temperature from March (black cross), April (red square), May (green triangle) and June (blue diamond) 2001: (a) α_g and (b) β_g with resampled error bars (plotted points are staggered around each forecast time for clarity).

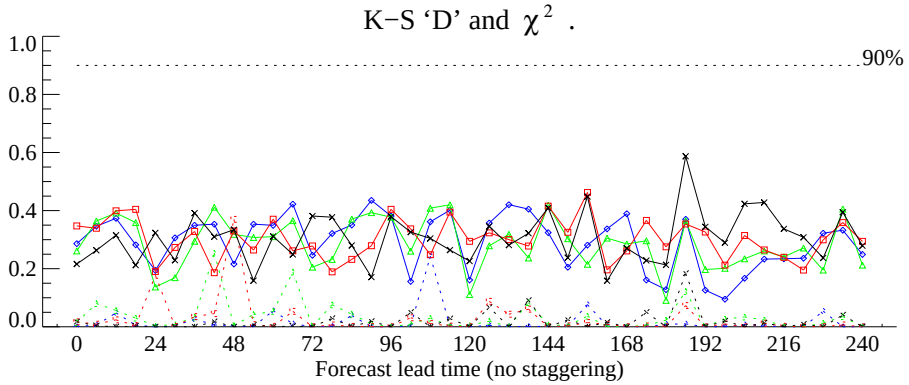

Figure 9: Statistical confidence information from the quasi-operational calibration system for upper 2m temperature from March (black cross), April (red square), May (green triangle) and June (blue circle) 2001: D (solid) and χ^2 (dashed) percentiles (see text for explanation).

Figure 9 shows time series' of the 'goodness of fit' parameters which enable the reliability of the fitted tails to be monitored, with the 90% 'warning level' highlighted by the horizontal dotted line. The solid lines show the percentile position of the global fit D within its bootstrapped distribution. Provided the resampling is done using a representative sample size from a sufficiently large data set this parameter shows how representative is the global fit of the 'climate' (as simulated by the resampling). Small percentage points indicate relatively small differences between the global fit and global histogram, compared with the resampled distributions, and therefore a good fit. Conversely, a 'large' percentage value (*i.e.* the 90% rejection level) indicates that the global fit is found towards the upper end of the resampled D distribution. Such a large difference between the global histogram and fit would not be expected to occur by chance, indicating the global fit is a poor representation of the 'climate' and so is unsuitable for forecasting that particular distribution tail. Figure 7(c) shows a case where the global fit lies within the bulk of the resampled distribution of D which indicates a consistent fit.

The dashed lines in figure 9 show the position within the χ^2 distribution (for the appropriate degrees of freedom taken from the global histogram) of the χ^2 value from the global fit. Provided that the assumption of normally-distributed errors on which the standard 'χ^2 test' is based holds then the percentage plotted roughly corresponds with the probability that the global χ^2 value would *not* arise by chance. The rejection level of 90% corresponds with the horizontal dotted line, beyond which the global fit is rejected.

In practice poor fits to the distribution arise due to insufficient outlier data. This occurs when too few data are pooled into the global fit or non- independent data are 'thinned out' by the imposed 24h. separation in time. Provided that sufficient months are included there is usually sufficient data (3 months from each of 2 years has been sufficient for the test set of 30 UK stations), however when the outlier weights improve towards the ideal value of 1/52 (as discussed above) the population of the outliers decreases. When this happens it would be necessary to pool many months' data to characterise the outlier tail distributions, which would lose sensitivity to model and/or seasonal changes. Fortunately this is unnecessary since small, near ideal outlier weight means that the outlier bin is no-longer overemphasised in the PDF and the effect of the fitted tail becomes insignificant. It is then straightforward, as well as being scientifically justified, to

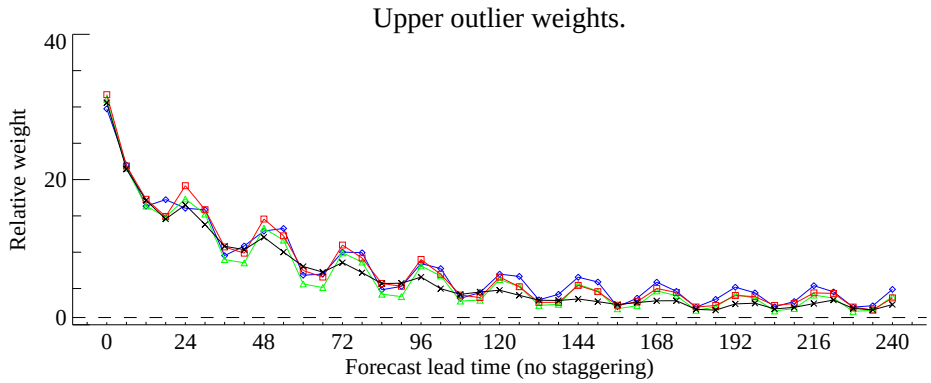

Figure 10: Upper outlier weights (normalised with the ideal weight) from the quasi-operational calibration system for upper 2m temperature from March (black cross), April (red square), May (green triangle) and June (blue circle) 2001.

replace the missing Weibull tails with a simple linear extrapolation since the probabilistic weight being extrapolated is only $\sim 2\%$.

The evolution of the upper outlier weights is illustrated in fig. 10 for 2m temperature, with respect to the 'ideal' relative weight of 1 indicated by the dashed horizontal line (which corresponds to the ideal weight of 1/52). There is a diurnal cycle in the weights, suggesting more outliers occur during the day than at night, and a greater diurnal variability is found in later months, perhaps due to the EPS not capturing the increasing observed daytime temperatures (which was also not corrected by the Kalman filter). Finally the weights decrease towards the (relative) limit of 1 with forecast time, following the increase in EPS spread.

5 Verification

Mylne et al. (2001) discussed the details of the operational verification of the new Previn system. In summary, they found that the Kalman filter and calibration successfully improved the probability forecasts for many events, even when the site-specific corrections were poor (such as with small amounts of precipitation). However the calibration was found to degrade forecasts of some extreme events (*e.g.* large amounts of precipitation), which was presumed to be due to the use of statistics derived from samples dominated by non-extreme events. Here the performance of the calibration system alone is illustrated using data from 2000/1. This analysis uses non-operational verification statistics which provides the opportunity to assess the statistical confidence that may be attached to the results.

5.1 95% confidence temperature ranges

Part of the remit of the Previn development project was to improve forecasts of 'first guess' forecasts of 95-percentile temperatures from the EPS. Define $T^{+}_{2.5\%}$ and $T^{-}_{2.5\%}$ such that (from eq. 5) $P\left(T < T^{-}_{2.5\%}\right) = 0.025$ and $P\left(T > T^{+}_{2.5\%}\right) = 0.025$, that is the forecast temperatures

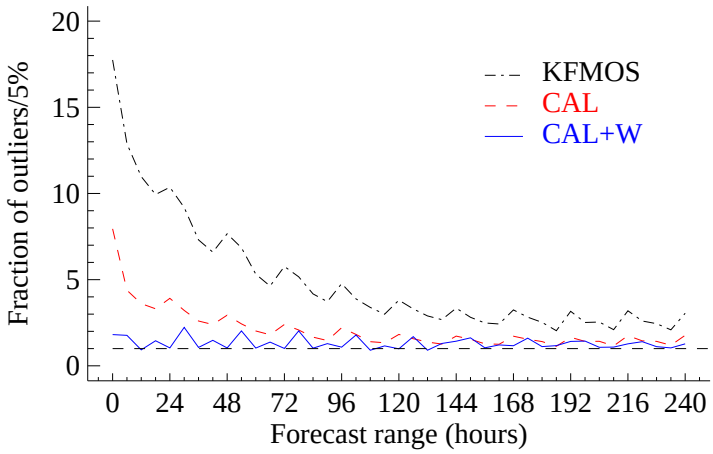

Figure 11: Verifying temperatures outside 95-percentile forecasts from December 2000 to June 2001. The three lines use the uncalibrated, Kalman-filter data labelled as 'KFMOS', calibrated probabilities as 'CAL' (without fitted Weibull tails) and calibrated probabilities with fitted Weibull tails as 'CAL+W'.

between which the observations should lie 95% of the time, *i.e.* $P\left(T^-_{2.5\%} \leq T \leq T^+_{2.5\%}\right) = 0.95$. Thus the problem becomes one of finding $T^+_{2.5\%}$ and $T^-_{2.5\%}$ from eq. 5

When the temperature outlier weights are large (as they are in fig 1 at T+12 and T+0–120 in fig. 10) the 95% limits on temperature fall mainly within the outlier bins. Under such circumstances it is the fitted Weibull tails that determine the 95% temperatures so this is a good test of both how the calibration improves the reliability of the probability forecasts and how well the fitted Weibull tails characterise the outlier distributions.

Figure 11 shows the frequency with which the verifying temperature lies *outside* the 95% limit divided by the 'ideal' of 5%, which is indicated by a horizontal dashed line. The figure is based on verification data from December 2000 to June 2001 (the most recent Previn data at the time of writing) giving an average of more than 3300 cases for each forecast time from 30 UK stations, where points were only included when both the calibrated weights and fitted tails were available. Evidently the uncalibrated temperature range (black, dash-dot line for 'KFMOS') is too narrow so that observed temperatures are found outside the 95% limits far too often. The error decreases with forecast range as the EPS spread improves but rarely falls below 3. Calibration using the weights alone (red, dashed line for 'CAL') improves the forecasts so that they are almost ideal from T+96–240h. Adding the fitted Weibull tails improves matters further where the outlier weights are largest so that the fully calibrated 95% temperature forecasts verify well at all forecast ranges (blue, solid line for 'CAL+W').

5.2 Brier skill scores

A traditional measure of probabilistic skill is the Brier score, defined as

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2 \tag{6}$$

where y_k and o_k are the forecast and observations, respectively, (see Wilks 1995, pp. 259–260 for details) and is analogous to a mean-squared probability error. To compare Brier scores between systems the *skill score* may be calculated

$$BSS = \frac{BS_{\text{ref}} - BS}{BS_{\text{ref}}}$$

which takes positive values for improved Brier skill, BS, compared to the reference system, BS_{ref} (since the Brier score is non-negative with a 'perfect' score being zero). Here the BSS is used to assess whether or not the calibration system improves over the Kalman filter, which is tacitly assumed to be beneficial, or skilful, with respect to the uncorrected EPS.

5.2.1 Temperature

Figure 12 shows Brier skill scores (BSS) for two temperature events which are (a) $P(T < -2°\text{C})$ and (b) $P(T < 10°\text{C})$. The error bars at each point are derived by randomly resampling the days from which the time series data used to calculate the individual BS for each of the three systems (KFMOS, CAL, CAL+W). Thus it is the skill score that is resampled (rather than randomly sampling each BS separately) with each bar indicating the 95% limits from 10,000 such samples. Therefore (assuming the resampling process is representative) the probability that the skill at a given forecast time would lie outside the limits of the error bars is 5%.

Figure 12(a) represents a relatively rare event ($T < -2°\text{C}$) for which the number of events occurring is small – the missing points from the figure correspond with forecast times at which there were no events verified (mainly daytime forecasts). The error bars are quite large and they indicate that it is not possible to say with confidence that either CAL+W or CAL are more skilful than KFMOS, except at T+18 and T+42 where the BSS and the lower 95% limit lie above zero. However, conversely, it is also not possible to conclude with confidence that either CAL+W or CAL are *less* skilful than KFMOS. The addition of the Weibull tails in this case improves the skill slightly, but not significantly (according to the resampling). Slightly better skill is seen for $T < 0°\text{C}$ (figure not shown) but with similar conclusions.

Figure 12(b) represents a more common event ($T < 10°\text{C}$) where the resampled error bars are smaller, indicating that the EPS produces a good forecast more often (more of the resampled scores are closer to the skill from the whole dataset). In this case the calibration produces significant improvements in Brier skill up to T+72 (corresponding with the largest weights), although including the Weibull tails (CAL+W) produces no significant improvement over CAL. Again, the calibration does not degrade the skill beyond the 95% limit (with respect to the Kalman filter) at any forecast time. Similar behaviour is found for forecasts of $T < 2°\text{C}$ (not shown) and $T < 5°\text{C}$, the plot for which is almost identical to $P(T > 5°\text{C})$ in fig. 13(a) (since changing from $P(T < 5°\text{C})$ to $P(T \geq 5°\text{C})$ leaves eq. 6 unchanged). The diurnal character of

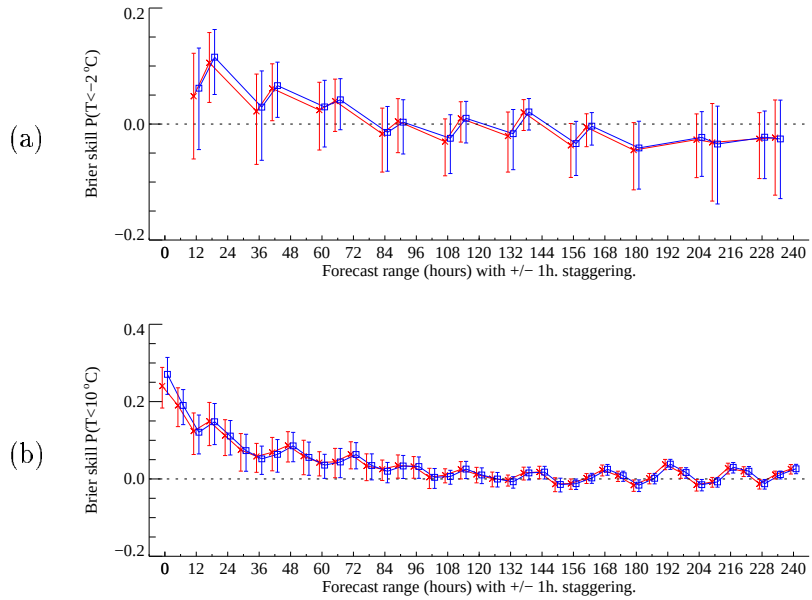

Figure 12: Brier skill scores from temperature probabilities for (a) $P(T < -2°\text{C})$ and (b) $P(T < 10°\text{C})$. Skill is calibrated (CAL+W, blue) or calibrated with no Weibull tails (CAL, red) against uncalibrated (KFMOS). Data were taken from January to May 2001 and have been displaced by $\pm$1h. about their nominal forecast time for clarity.

the temperature skill scores is evident, with slightly better skill compared to KFMOS at T+48, 72, 168, 192, 216 and 240, all of which verify at midday.

Figures 13(a) and (b) show similar graphs of Brier skill to fig. 12 for the probability of forecast temperatures exceeding a threshold. Figure 13(b) is for a relatively rare event ($P(T > 20°\text{C})$ – *c.f.* fig. 12(a) above) where only the data verifying at 12/18Z are plotted. The overall behaviour of both graphs is similar, with a statistically-significant improvement in skill (according to the resampling) early in the forecast and unimproved skill later, other than two points in fig. 13(a) where the calibration skill is slightly less than uncalibrated KFMOS (T+222 and T+240h). Thresholds between 5°C and 20°C exhibit similar patterns, such as $P(T > 10°\text{C})$ for which the figure is not shown here as it is similar to fig. 12(b) for $P(T < 10°\text{C})$.

5.2.2 Precipitation

Figures 14(a) and (b) show plots of calibration skill for 12 hour precipitation accumulations and these are markedly different. In fig. 14(a) the skill of forecasts of the probability of 'a trace of rain' are shown to be significantly better with calibration than without at all forecast ranges. There is a very strong 'diurnal' variation which may be due to the bias correction used in place of the Kalman filter. Corrections are made at 0Z, 6Z, 12Z and 18Z for all forecast ranges. As the EPS starts at 12Z the first 12h. PPT accumulation occurs at 0Z based on the first 12h. of the forecast, during which time the precipitation from the convective parametrisation is 'spinning up'. Thus the KFMOS correction is poor and the calibration weights are very large (as shown

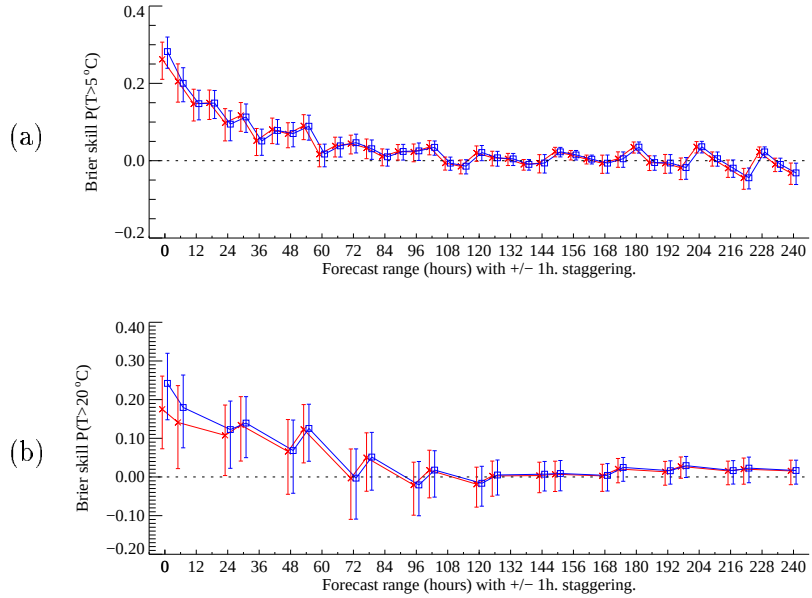

Figure 13: As in figure 12 except showing Brier skill scores from temperature probabilities for (a) $P(T > 5°\text{C})$ and (b) $P(T > 20°\text{C})$ with $\pm$1h. staggering.

in 15(b)) for all 0Z verifying times (T+12, 36, ..., 228h.) This results in the observed increased skill for the calibration with respect to the KFMOS. The skill remains positive throughout the 10 day forecast and there is no apparent change due to the Weibull tails. This is unsurprising as in this case the low PPT threshold will lie near to, if not inside the lower outlier bin for which no Weibull tail is calculated (only an upper tail is generated for precipitation).

In fig. 14(b) a climatologically rarer event is considered, that of the probability of more than 10mm rain in 12 hours. In this figure there is evidence that the calibration degrades the forecast skill slightly for forecasts of $PPT12 > 10$mm/12h at long forecast ranges, with respect to the uncalibrated KFMOS forecast (this conclusion was also drawn in Mylne et al. (2001) using a different verification process). In fact the upper outlier weights for PPT12 are quite close to 1/52 beyond T+48 as shown in figure 15(a) (the relative weight is less than 3) whilst the lower outlier weights in figure 15(b) remain large. The large lower outlier weight means that for a given sampling period (3 months for the weights; 6 months for the tails in this case) more of the data lie in (or close to) the lower outlier bin, relatively less in the other bins. Therefore the bins with small weights are less well sampled than those with large weights and in the case of PPT12 the upper outlier bin is the least well sampled of them all, which is likely to have a detrimental effect on forecasts of large precipitation thresholds (towards the right of the PDF).

Since the bias correction is not particularly successful in this case, different rainfall climatologies will remain within the station groups, which is not the case for Kalman-filtered wind speed or temperature, as discussed previously (*c.f.* figures 3 (b) and (c)). Thus the 'average' Weibull fit is will not be representative of individual stations. Alternative strategies for improving precipitation forecasts will need to be investigated.

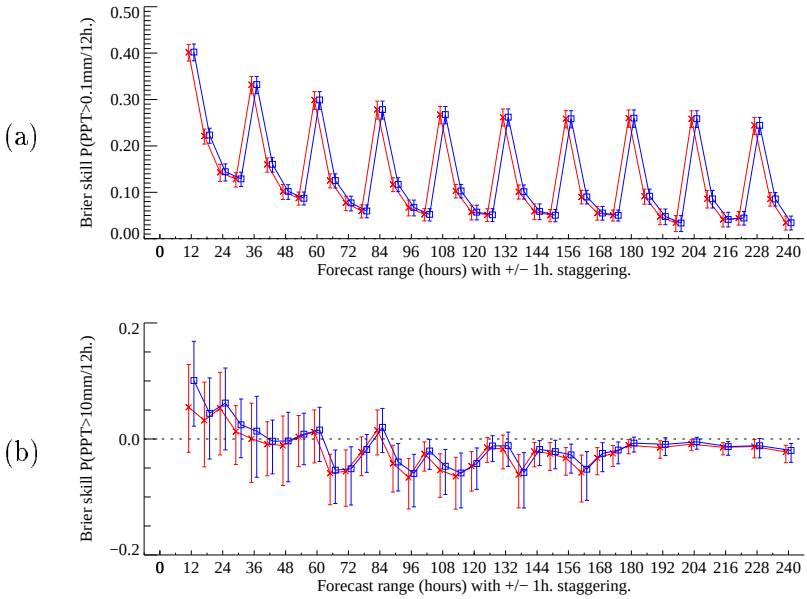

Figure 14: As in figure 12 except showing Brier skill scores from precipitation probabilities for (a) $P(\text{PPT} > 0.1\text{mm/12h.})$ and (b) $P(\text{PPT} > 10\text{mm/12h.})$ with $\pm$1h. staggering.

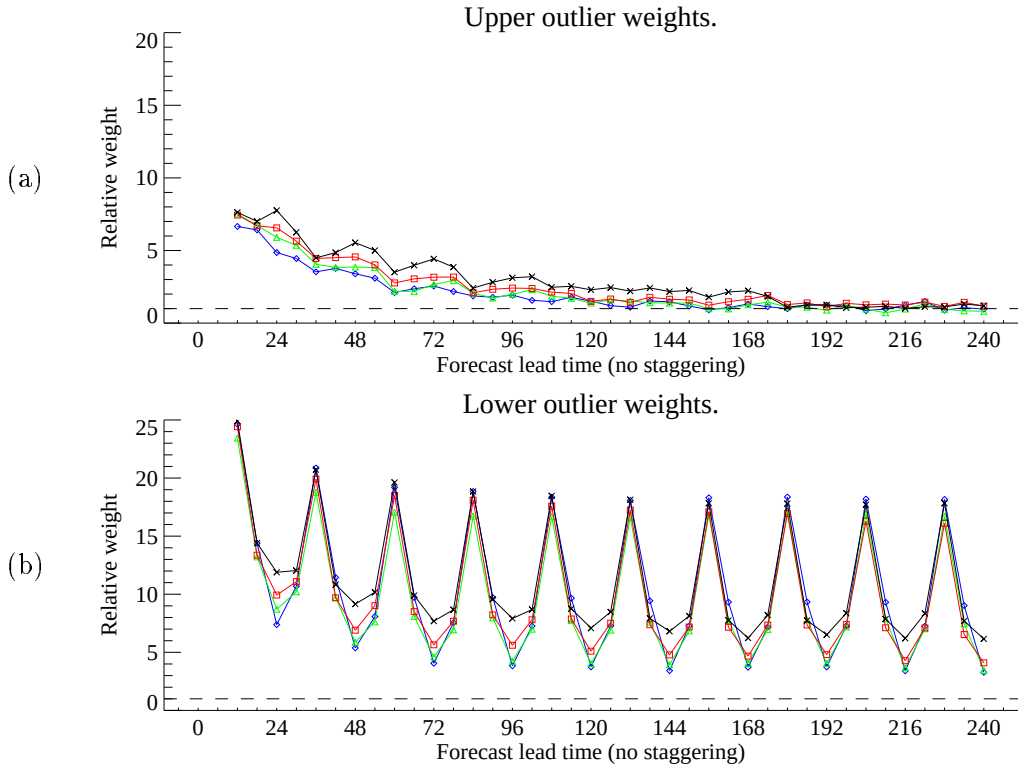

Figure 15: (a) Upper and (b) lower outlier weights for 12h. precipitation (PPT12) from the quasi-operational calibration system for March (black cross), April (red square), May (green triangle) and June (blue circle) 2001.

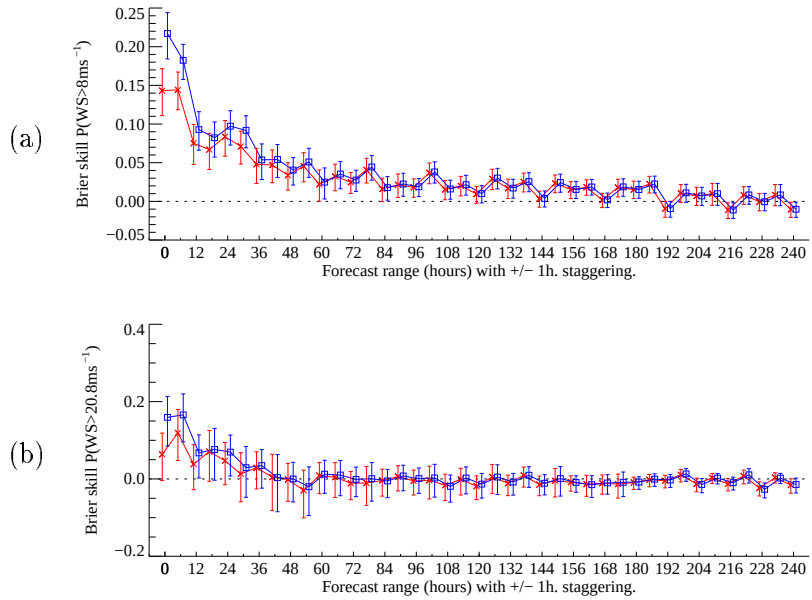

Figure 16: As in figure 12 except showing Brier skill scores for wind speed probabilities for (a) P(10m wind $\geq$ 8ms^{-1}) and (b) P(10m wind $\geq$ 20.8ms^{-1}) with $\pm$1h. staggering. (**NB:** Beaufort Force 5 and 9 are equivalent to 8–10.7ms^{-1} and 20.8–24.4ms^{-1}, respectively.)

5.2.3 Wind speed

Finally, in figure 16(a) and (b) are plots of Brier skill for two wind speed thresholds. For the relatively common event in fig. 16(a) (10m wind speed $>$ 8ms^{-1}) there is a significant increase in skill due to the calibration over the Kalman filter up to 5 days ahead (and beyond for data verifying at times other than 12Z). There is a smooth transition between the skill shown in fig. 16(a) and fig. 16(b) (10m wind speed $>$ 20.8ms^{-1}), although intermediate thresholds are not shown here. At the higher wind speeds there is little improvement in skill over the Kalman filter but, conversely, no significant detriment, either. Interestingly, on fig. 16(a) at T+0 (and almost T+6) the lower outlier weights are so large that there is a significant benefit to using the fitted Weibull tails in addition to the calibrated weights.

6 Conclusion

The calibration system, including Weibull tails, has clear benefits for site-specific probabilistic forecasts, particularly at early forecast times (where the EPS is known to spread insufficiently and the outlier weights are large). It is encouraging that the calibration does not degrade forecast skill at long forecast ranges, where the weights are low (and so the tails are less well sampled), except for the larger precipitation thresholds. In the latter case improved skill might be achieved by improving the station grouping and/or the sample sizes (in a consistent fashion) but the current strategy (constrained by data availability) means that the calibration is slightly detrimental to forecast skill (compared with the uncalibrated KFMOS) beyond T+72 for PPT12

thresholds of 10mm or above and should not be used for such forecasts.

Evidently the addition of Weibull tails is beneficial for specific applications (such as 95% confidence temperatures) but otherwise has only a small effect on Brier skill scores. However it is known that improving the Brier skill score does not necessarily equate with increased measures of skill such as 'forecast value' (*e.g.* Mylne 2001). Verification of Previn with the Kalman filter and calibration system based on 'value' is in preparation which will prove to be a significant test of the system and have relevance to customer applications.

References

Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Climatol.* **9**: 1518–1530.

Erickson, D. J. and Taylor, J. A. (1989). Non-Weibull behavior observed in a model-generated global surface wind field frequency distribution, *J. Geophys. Res.* **94**: 12693–12698.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Wea. Rev.* **129**: 550–560.

Hamill, T. M. and Colucci, S. J. (1997). Verification of eta-rsm short-range ensemble forecasts, *Mon. Wea. Rev.* **125**: 1312–1327.

Legg, T. P., Mylne, K. R. and Woolcock, C. (2001). The use of medium-range ensembles at the Met Office. I: PREVIN - a system for the production of probabilistic forecast information from the ECMWF EPS, *Technical report*, The Met Office. Forecasting Research Technical Note No. 333.

Mylne, K. R. (2001). Decision-making from probability forecasts using calculations of forecast value, *Technical report*, The Met Office. Forecasting Research Tech Note No 335.

Mylne, K. R., Darvell, R., Denholm-Price, J. C. W. and Woolcock, C. (2001). Previn development project - verification report. Unpublished Met Office report available internally at `http://www-nwp/~frkm/FSSSI_Previn/VerifReport.doc`.

Press, W. H., , Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical recipes: The art of scientific computing*, Cambridge University Press.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences - An Introduction*, Vol. 59 of *International Geophysics Series*, Academic Press.

Woolcock, C. (2001). Kalman filter experiments for ensembles. Unpublished Met Office report available internally at `http://www-nwp/~frkm/FSSSI_Previn/Kalman_Filter_experiments_for_Ensembl%es.doc`.